Bayesian Joint Modelling for Object Localisation in Weakly Labelled Images

Zhiyuan Shi, Timothy M. Hospedales, Tao Xiang

Abstract—We address the problem of localisation of objects as bounding boxes in images and videos with weak labels. This weakly supervised object localisation problem has been tackled in the past using discriminative models where each object class is localised independently from other classes. In this paper, a novel framework based on Bayesian joint topic modelling is proposed, which differs significantly from the existing ones in that: (1) All foreground object classes are modelled jointly in a single generative model that encodes multiple object co-existence so that "explaining away" inference can resolve ambiguity and lead to better learning and localisation. (2) Image backgrounds are shared across classes to better learn varying surroundings and "push out" objects of interest. (3) Our model can be learned with a mixture of weakly labelled and unlabelled data, allowing the large volume of unlabelled images on the Internet to be exploited for learning. Moreover, the Bayesian formulation enables the exploitation of various types of prior knowledge to compensate for the limited supervision offered by weakly labelled data, as well as Bayesian domain adaptation for transfer learning. Extensive experiments on the PASCAL VOC, ImageNet and YouTube-Object videos datasets demonstrate the effectiveness of our Bayesian joint model for weakly supervised object localisation.

Index Terms—Object Detection, Topic Modelling, Weakly Supervised Learning, Bayesian Domain Transfer, Probabilistic Modelling.

1 Introduction

Object recognition is a challenging problem especially at a large scale because of variabilities in object appearance, viewpoint, illumination and pose [1], [2], [3]. Fully/strongly annotated data is thus typically required to learn a generalisable model for tasks such as object classification [4], detection [5], [6], and segmentation [3], [7], [8]. In fully annotated images, such as those in the PASCAL VOC object classification or detection challenges [9], not only the presence of objects, but also their locations are labelled, typically in the form of bounding boxes. Such a strong manual annotation of objects is time-consuming and laborious. Consequently, although media data is increasingly available with the prevalence of sharing websites such as Flickr, the lack of annotated images, particularly strongly annotated ones, becomes the new barrier that prevents tasks such as object detection from scaling to thousands of classes [10].

One approach to this challenge is weakly supervised object localisation (WSOL): simultaneously locating objects in images and learning their appearance using only weak labels indicating presence/absence of the objects of interest. The WSOL problem has been tackled using various approaches [11], [4], [12], [13], [10], [14], [15]. Most of them address the task as a weakly supervised learning problem, particularly as a multi-instance learning (MIL) problem, where images are bags, and potential object locations are instances. These methods are typically discriminative in nature and attempt to localise each class of objects independently from the other classes. However, localising objects of different classes independently has a number of limitations: (1) It fails to exploit the knowledge that different objects often co-exist within an image (see Fig. 1). For instance, knowing that some images have both a horse and a person, in conjunction with a joint model for both classes – the person can be "explained away" to reduce ambiguity about the horse's appearance, and vice versa. Ignoring this increases ambiguity for each class. (2) Although object classes vary in appearance, the background appearance is relevant to them all (e.g. sky, tree, and grass are constant features of an image regardless of the foreground object classes). When different classes are modelled independently, the background must be re-learned repeatedly for each class, when it would be more statistically robust [16] to share this common knowledge.

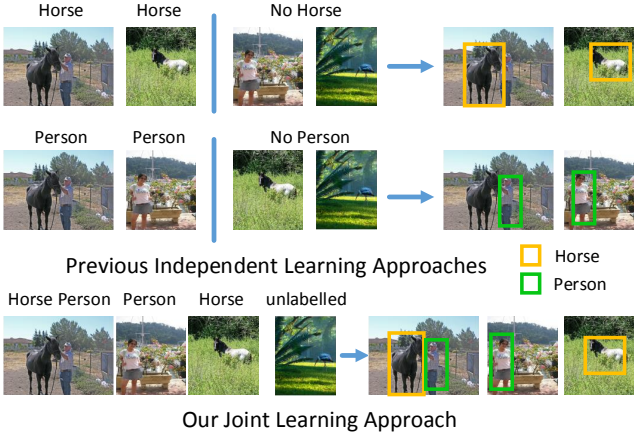


Fig. 1: Different types of objects often co-exist in a single image. Our joint learning approach differs from previous approaches which localise each object class independently.

In this paper, a novel framework based on Bayesian

latent topic models is proposed to overcome the mentioned limitations. In our framework, both multiple object classes and background types are modelled jointly in a single generative model as latent topics, in order to explicitly exploit their co-existence relationship (see Fig. 1). As bag-of-words (BoW) models, conventional latent topic models have no notion of localisation. We overcome this problem by incorporating an explicit notion of object location.

Our generative model based framework has the following advantages over previous discriminative approaches:

Joint vs. independent modelling By jointly modelling different classes of objects and background, our model is able to exploit multiple object co-occurrence, so each object known to appear in an image can help disambiguate the location of the others by accounting for some of the pixels. This is illustrated by the left column of Fig. 1, where modelling horse and person jointly helps the localisation of both objects since each single pixel can only be explained by one object, not both. Meanwhile, a single set of shared background topics are learned once for all object classes. This is due to the nature of a generative model – every pixel in the image must be accounted for. Even though learning background appearance can further disambiguate the location of objects, this appears to be an extremely hard task given that no labels are provided regarding background (people tend to focus on the foreground when annotating an image). However, by learning them jointly with the foreground objects and using all training images available, this task can be fulfilled effectively by the proposed model.

Integration of prior knowledge Exploiting prior knowledge or top-down cues about appearance or geometry (e.g., position, size, aspect ratio) should be supported if available to offset the weak labels. Our framework is able to incorporate, when available, prior knowledge about appearances of objects in a more systematic way as a Bayesian prior. Specifically, we exploit the prior intuition that objects are spatially compact relative to the background. We can also optionally exploit external human or internal data-driven prior about typical object size, location and appearance as a Bayesian prior. Going beyond within-class priors, we also show that cross-class appearance similarity can be exploited. For instance, the model can exploit the fact that "bike" is more similar to "motorbike" than "aeroplane".

Bayesian domain adaptation A central challenge for building generally useful recognition models is providing the capability to adapt models trained on one domain or dataset to new domains or datasets [17]. This is important because any given domain or dataset is intentionally or unintentionally biased [18], so transferring models directly across domains generally performs poorly [18]. However, with appropriate adaptation, source and target domain data can be combined to out-perform target domain data alone [17]. We can leverage our model's Bayesian formulation to provide domain adaptation in a WSOL context.

Semi-supervised learning Since there are effectively unlimited quantity of unlabelled data available on the Internet (compared to limited quantity of manually annotated data), a valuable capability is to exploit this existing unlabelled data in conjunction with limited weakly labelled data to improve learning. As a generative model, our framework is naturally suited for semi-supervised learning (SSL). Unlabelled data are included and the label variables for these instances left unclamped (i.e. no supervision is enforced). Importantly, unlike conventional SSL approaches [19], our model does not require that all the unlabelled data are instances of known classes, making it more applicable to realistic SSL applications.

2 Related Work

Weakly supervised object localisation Weakly supervised learning (WSL) has attracted increasing attention as the volume of data which we are interested in learning from grows much faster than available annotations. Weakly supervised object localisation (WSOL) is of particular interest [11], [20], [12], [15], [13], [21], [4], [22], [15], [23], due to the onerous demands of annotating object location information. Many studies [4], [11] have approached this task as a multi-instance learning [24], [25] problem. However, only relatively recently have localisation models capable of learning from challenging data such as the PASCAL VOC 2007 dataset been proposed [11], [20], [12], [13], [21]. Such data is especially challenging because objects may occupy only a small proportion of an image, and multiple objects may occur in each image: corresponding to a multi-instance multi-label problem [26]. Three types of cues are exploited in existing WSL object localisation approaches: (1) *saliency* – a region containing an object should look different from the majority of (background) regions. The object saliency model in [27] is widely used in most recent work [11], [12], [10], [20], [17] as a preprocessing step to propose a set of candidate object locations so that the subsequent computation is reduced to a tractable level, (2) *intra-class* – a region containing an object should look similar to the regions containing the same class of objects in other images [20], and (3) *inter-class* – the region should look dissimilar to any regions that are known to not contain the object of interest [11], [12], [13]. One of the first studies to combine the three cues for WSOL was [11] which employed a conditional random field (CRF) and generic prior object knowledge learned from a fully annotated dataset. Later, [13] presented a solution exploiting latent SVMs. Recent studies have explicitly examined the role of intra- and inter-class cues [20], [12], as well as transfer learning [21], [10], for this task. Similar to the above approaches for weakly labelled images, [28], [17] proposed video based frameworks to deal with motion segmented tubes instead of bounding-boxes. In contrast to these studies, which are all based on discriminative models, we introduce a generative topic model based approach that exploits all three cues, as well as joint multi-label, semi-supervised and cross-domain adaptive learning.

Exploiting prior knowledge Prior knowledge has been exploited in existing WSOL works [11], [12], [13]. Recognition or detection priors can be broadly broken down into appearance and geometry (location, size, aspect) cues, and

can be provided manually, or estimated from data. The most common use has been crude: to generate candidate object locations based on a pre-trained model for generic objectness [29], i.e. the previously mentioned saliency cue. This reduces the search space for discriminative models. Beyond this, geometry priors have also been estimated during learning [11]. We can not only exploit such simple appearance and geometry cues as model priors, but also go beyond to exploit a richer object hierarchy, which has been widely exploited in classification [30], [31], [32], [16] and to a less extent detection [10], [7]. More specifically, we leverage WordNet, a large lexical database based on linguistics [33]. WordNet provides a measure of prior appearance similarity/correlation between classes, and we use this prior to regularise appearance learning. Such cross-class appearance correlation information is harder to use in previous WSOL approaches because different classes are trained separately. Interestingly, our model uniquely shows positive results for WordNet-based appearance correlation (see Sec. 8.2), in contrast to some recent studies [32], [16] that found no or limited benefit from exploiting WordNet based cross-class appearance correlation for recognition. Compared to the classification task, this inter-class correlation information is more valuable for WSOL because the task is more ambiguous. Specifically, the interdependent localisation and appearance learning aspects of the task adds an extra layer of ambiguity – the model might be able to learn the appearance if it knew the location, but it will never find the location without knowing appearance. Our work is related to [10] where hierarchical cross-class appearance similarity is used to help weakly supervised object localisation in ImageNet by transfer learning. However, a source dataset of fully annotated images are required in their work, whilst our model exploits the correlation directly for the target data which is only weakly labelled.

Cross domain/dataset learning Domain adaptation [34] methods aim to exploit prior knowledge from a source domain/dataset to improve the performance and/or reduce the amount of annotation required in a target domain/dataset (see [35] for a review). Many conventional approaches are based on SVMs for which the target domain can be considered a perturbed version of the source domain, and thus learning proceeds in the target domain by regularising it toward the source [36]. More recently, transductive SVM [37], Multiple Kernel Learning (MKL) [38], and instance constraints [39] have been exploited. In contrast to these discriminative approaches, we exploit a simple and efficient Bayesian adaptation approach similar in spirit to [34], [40]. Posterior parameters from the source domain are transferred as priors for the target, which are then adapted based on observed target domain data via Bayesian learning. Going beyond simple within-modality dataset bias, recent studies [17], [28] have adapted object detectors from video to image or reverse. We show that our approach can achieve the image-video domain transfer within a single framework.

Exploiting unlabelled data Semi-supervised learning [19] methods aim to reduce labelling requirements and/or improve results compared to only using labelled data. Most existing SSL approaches assume a training set with a mix of fully labelled and weak or unlabelled [41], [10] data, while we use weak and unlabelled data alone. The existing (discriminative) line of work focusing on WSOL [11], [13], [42], [43] has not generally exploited unlabelled data, and cannot straightforwardly do so.

Topic models for image understanding Latent topic models (LTMs) were originally developed for unsupervised text analysis [44], and have been successfully adapted to both unsupervised [45], [46] and supervised image understanding problems [47], [48], [49], [50], [51]. Most studies have addressed the simpler tasks of classification [50], [51] and annotation [50], [52]. Our model differs from the existing ones in two main aspects: (i) Conventional topic models have no explicit notion of the spatial location and extent of an object in an image. This is addressed in our model by modelling the spatial distribution of each topic. Note that some topic model based methods [49], [48] can also be applied to object localisation. However, the spatial location is obtained from a pre-segmentation step rather than being explicitly modelled. (ii) The other difference is more subtle – existing supervised topic models such as CorrLDA [52], SLDA [50] and derivatives [49] only weakly influence the learned topics. This is because the objective is the sum of visual words and label likelihoods, and visual words vastly outnumber annotations, thus dominating the result [51]. The limitation is serious for WSOL as the labels are already weak and they must be used to their full strength. In this work, a learning algorithm with topic constraints similarly to [2] is formulated to provide stronger supervision which is demonstrated to be much more effective than the conventional supervised topic models in our experiments (see supplementary material). With these limitations addressed, we can exploit the potential of a generative model for domain adaptation, joint-learning of multiple objects and semi-supervised learning.

Other joint learning approaches An approach similar in spirit to ours in the sense of jointly learning a model for all classes is that of Cabral *et al* [53]. This study formulates multi-label image classification as a matrix completion problem, which is also related to our factoring images into a mixture of topics. However we add two key components of (i) a stronger notion of the spatial location and extent of each object, and (ii) the ability to encode human knowledge or transferred knowledge through a Bayesian prior. As a result, we are able to address more challenging data than [53] such PASCAL VOC. Multi-instance multi-label (MIML) [26] approaches provide a mechanism to jointly learn a model for all classes [54], [55]. However, because these methods must search for a discrete space (of positive instance subsets), their optimisation problem is harder than the smooth probabilistic optimisation here. Finally, while more elaborate joint generative learning methods [56], [49] exist, they are more complicated than necessary for WSOL and do not scale to the size of data required here.

Feature fusion Combining multiple complementary cues has been shown to improve classification performance in object recognition [57], [58], [59], [38]. Two simple feature

fusion methods have been widely used in existing work: early fusion which combines low-level features [60] early (feature concatenation) and late (score level) fusion [11], [20]. Multiple kernel learning (MKL) approaches have attracted attention as a principled mid-level approach to combining features [59], [58]. Similarly to MKL, our framework provides a principled and jointly-learned mid-level probabilistic fusion via its generative process.

Contributions In summary, this paper makes the following contributions: (1) We propose the novel concept of joint modelling of all object classes and background for weakly supervised object localisation. (2) We formulate a novel Bayesian topic model suitable for object localisation, which can use various types of prior knowledge including an inter-category appearance similarity prior. (3) Our Bayesian prior enables the model to easily borrow available domain knowledge from existing auxiliary datasets and adapt it to a target domain. (4) We further exploiting unlabelled data for improving weakly supervised object localisation. (5) Extensive experiments on the PASCAL VOC 2007 [9] and ImageNet [61] show that our model surpasses existing competitors and achieves state-of-the-art performance. A preliminary version of our work was described in [62].

3 Joint Topic Model for Objects and Background

In this section, we introduce our new latent topic model (LTM) [44] approach to the weakly-supervised object localisation task. Applied to images, conventional LTMs factor images into combinations of latent topics [45], [46]. Without supervision, these topics may or may not correspond to anything of semantic relevance to humans. To address the WSOL task, we need to learn what is unique to all images sharing a particular label (object class), while explaining away both the pixels corresponding to other annotated objects as well as other shared visual aspects (background) which are irrelevant to the annotations of interest. We will achieve this in a LTM framework by applying weak supervision to partially constrain the available topics for each image. This constraint is enforced by label/topic clamping to ensure that each foreground topic corresponds to an object class of interest.

More specifically, to address the WSOL task, we will factor images into unique combinations of K shared topics. If there are C classes of objects to be localised, $K^{fg} = C$ of these will represent the (foreground) classes, and $K^{bg} = K - K^{fg}$ topics will model background data to be explained away. Each topic thus corresponds to one object class or one type of background. Let T^{fg} and T^{bg} index foreground and background topics respectively. An image is represented using a Bag-of-Words (BoW) representation for each of $f = 1 \dots F$ different types of features (see Sec. 8.1 for the specific appearance features used). After learning, each latent topic will encode both a distribution over the V_f sized appearance vocabulary of each feature f and also over the spatial location of these words within each image. Formally, given a set of J training images, each labelled with any number of the C foreground classes, and represented as bags of words $\mathbf{x}_{jf}$, the generative process of our model (Fig. 2) is as follows (notation is summarised in Table 1 for convenience):

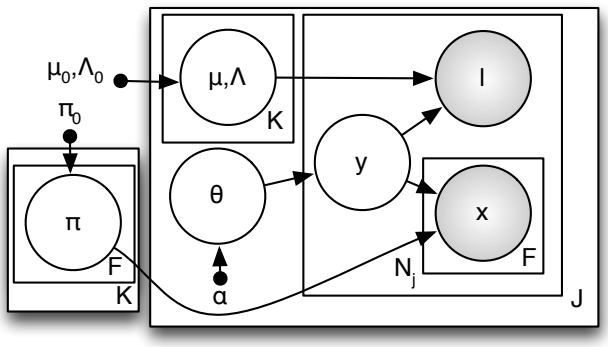


Fig. 2: Graphical model for our WSOL joint topic model. Shaded nodes are observed.

For each topic $k \in 1 \dots K$:

1) For each feature representation $f \in 1 \dots F$:
   a) Draw an appearance distribution $\boldsymbol{\pi}_{kf} \sim \text{Dir}(\boldsymbol{\pi}^0_{kf})$ following the Dirichlet distribution

For each image $j \in 1 \dots J$:

1) Draw foreground and background topic distribution $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha}_j)$, $\boldsymbol{\alpha}_j = [\boldsymbol{\alpha}^{fg}_j, \boldsymbol{\alpha}^{bg}_j]$ where the Dirichlet distribution parameter α_j reflects prior knowledge of the presence of each object class or background in the image j. Both $\boldsymbol{\theta}_j$ and $\boldsymbol{\alpha}_j$ are K dimensional.
2) For each foreground topic $k \in T^{fg}$ draw a location distribution: $\{\boldsymbol{\mu}_{kj}, \Lambda_{kj}\} \sim \mathcal{NW}(\boldsymbol{\mu}^0_k, \Lambda^0_k, \beta^0_k, \nu^0_k)$
3) For each observation (visual word) $i \in 1 \dots N_j$:
   a) Draw topic $y_{ij} \sim \text{Multi}(\boldsymbol{\theta}_j)$
   b) Draw a location:
      $\mathbf{l}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{y_{ijj}}, \Lambda^{-1}_{y_{ijj}})$ if $y_{ij} \in T^{fg}$ or
      $\mathbf{l}_{ij} \sim Uniform$ if $y_{ij} \in T^{bg}$
   c) For each feature representation $f \in 1 \dots F$:
      i) Draw visual word $x_{ijf} \sim \text{Multi}(\boldsymbol{\pi}_{y_{ij}f})$

where Multi, Dir, $\mathcal{N}$, $\mathcal{NW}$ and $Uniform$ respectively indicate Multinomial, Dirichlet, Normal, Normal-Wishart and uniform distributions with the specified parameters. These prior distributions are chosen mainly because they are conjugate to the word, topic and location distributions, and hence enable efficient inference. For the visual word spatial location, the foreground and background distributions are of different forms – normal for foreground and uniform for background. This is to reflect the intuition that foreground objects tend to be compact and background much less so. The joint distribution of all observed $O = \{\mathbf{x}_{jf}, \mathbf{l}_j\}^{J,F}_{j,f=1}$ and latent $H = \{\{\boldsymbol{\pi}_{kf}\}^{K,F}_{k,f=1}, \{\mathbf{y}_j, \boldsymbol{\mu}_{kj}, \Lambda_{kj}, \boldsymbol{\theta}_j\}^{K,J}_{k,j=1}\}$ variables given parameters $\Pi = \{\{\boldsymbol{\pi}^0_{kf}\}^{K,F}_{k,f=1}, \{\boldsymbol{\mu}^0_k, \Lambda^0_k, \beta^0_k, \nu^0_k\}^K_{k=1}, \{\boldsymbol{\alpha}_j\}^J_{j=1}\}$ in our

$x_{ijf} = 1...V_f$	Visual word i in image j for feature f
$\mathbf{l}_{ij}$	Location of visual word i in image j
$y_{ijk} = 1 \dots K$	Topic (object) for explaining visual word x_{ijf}
$\boldsymbol{\alpha}_j$	Annotation / topic prior for image j
$\boldsymbol{\theta}_j$	Dirichlet topic proportion in image j
$\boldsymbol{\pi}^0_{kf}$	Appearance prior for topic/class k in feature f
$\boldsymbol{\pi}_{kf}$	Dirichlet appearance for topic/class k in feature f
$\boldsymbol{\mu}^0_k, \Lambda^0_k$	$\mathcal{NW}$ Location prior for class k
$\boldsymbol{\mu}_{kj}, \Lambda^{-1}_{kj}$	Gaussian location of object class k in image j

TABLE 1: Summary of model variables and parameters

model is therefore:

$$p(O, H|\Pi) = \prod_k^K \prod_f^F p(\boldsymbol{\pi}_{kf}|\boldsymbol{\pi}^0_{kf}) \quad (1)$$

$$\cdot \prod_j^J p(\boldsymbol{\theta}_j|\boldsymbol{\alpha}_j) \left[\prod_k^K p(\boldsymbol{\mu}_{jk}, \Lambda_{jk}|\boldsymbol{\mu}^0_k, \Lambda^0_k, \beta^0_k, \nu^0_k) \right.$$

$$\left. \left(\prod_i^{N_j} p(\mathbf{l}_{ij}|\boldsymbol{\mu}_{jk}, \Lambda^{-1}_{jk}) \prod_f^F p(x_{ijf}|y_{ij}, \boldsymbol{\pi}_{y_{ij}f}) p(y_{ij}|\theta_j) \right) \right] \quad (2)$$

4 MODEL LEARNING

Inference via variational message passing Learning our model involves inferring the following quantities: the appearance of each object class for each feature type, $\boldsymbol{\pi}_{kf}, k \in T^{fg}$ and each background type, $\boldsymbol{\pi}_{kf}, k \in T^{bg}$ for each feature type f; the word-topic distribution (soft segmentation) of each image $\mathbf{z}_j$, the proportion of visual words (related to the proportion of pixels) in each image corresponding to each class or background $\boldsymbol{\theta}_j$, and the location of each object $\boldsymbol{\mu}_{jk}, \Lambda_{jk}$ in each image (mean and covariance of a Gaussian). To learn the model and localise all the weakly annotated objects, we wish to infer the posterior $p(H|O, \Pi) = p(\{\mathbf{y}_j, \boldsymbol{\mu}_{jk}, \Lambda_{jk}, \boldsymbol{\theta}_j\}^{K,J}_{k,j}, \{\boldsymbol{\pi}_{kf}\}^{K,F}_{k,f}|\{\mathbf{x}_{jf}, \mathbf{l}_j\}^{J,F}_{j=1,f=1}, \Pi)$. This is intractable to solve directly; however a variational message passing (VMP) [63] strategy can be used to obtain a factored approximation $q(H|O, \Pi)$ to the posterior:

$$q(H|O, \Pi) = \prod_{k,f} q(\boldsymbol{\pi}_{kf}) \prod_j q(\boldsymbol{\theta}_j) q(\boldsymbol{\mu}_{jk}, \Lambda_{jk}) \prod_i q(y_{ij}). \quad (3)$$

Under this approximation a VMP solution is obtained by deriving integrals of the form $\ln q(\mathbf{h}) = E_{H \setminus \mathbf{h}} [\ln p(H, O)] + K$ for each group of hidden variables $\mathbf{h}$, thus obtaining the following updates for the sufficient statistics (indicated by tilde) of each variable:

$$\tilde{\theta}_{jk} = \alpha_{jk} + \sum_i \tilde{y}_{ijk}, \quad (4)$$

$$\tilde{y}_{ijk} \propto \int_{\boldsymbol{\mu}_{jk}, \Lambda_{jk}} \mathcal{N}(\mathbf{l}_{ij}|\boldsymbol{\mu}_{jk}, \Lambda^{-1}_{jk}) q(\boldsymbol{\mu}_{jk}, \Lambda_{jk})$$

$$\cdot \prod_f^F \exp\left(\Psi(\tilde{\pi}_{x_{ijf} y_{ij} f}) - \Psi(\sum_v \tilde{\pi}_{v y_{ij} f}) \right)$$

$$\cdot \exp\left(\Psi(\tilde{\theta}_{j y_{ijk}}) \right), \quad (5)$$

$$\tilde{\pi}_{vkf} = \pi^0_{vkf} + \sum_{ij} \mathbf{I}(x_{ijf} = v) \tilde{y}_{ijk}, \quad (6)$$

where Ψ is the digamma function, $v = 1 \dots V_f$ ranges over the BoW appearance vocabulary, $\mathbf{I}$ is the indicator function which returns 1 if its argument is true, and the integral in the second line returns the student-t distribution over $\mathbf{l}_{ij}$, $\mathcal{S}(\mathbf{l}_{ij}|\tilde{\boldsymbol{\mu}}_{jk}, \tilde{\Lambda}^{-1}_{jk} P, \tilde{\beta}_{jk}, \tilde{\nu}_{jk})$. Within each image j, standard updates [64] apply for the sufficient statistics $\{\tilde{\boldsymbol{\mu}}_{jk}, \tilde{\Lambda}_{jk}, \tilde{\beta}_{jk}, \tilde{\nu}_{jk}\}$ of the Normal-Wishart parameter posterior $q(\boldsymbol{\mu}_{jk}, \Lambda_{jk})$. The update in Eq. (5) (estimating the object explaining each pixel) is the most non-standard for LTMs; this is because it contains a top-down contribution (the third term), and two bottom-up contributions from the location and appearance (the first and second terms respectively). The model is learned by iterating the updates of Eqs. (4)-(6) for all images j, words i, topics k and vocabulary v.

Supervision via label-topic constraints In conventional topic models, the α parameter encodes the expected proportion of words for each topic. In our weakly supervised topic model, we use α to encode the supervision from weak labels. In particular, we set α^{fg}_j as a binary vector with $\alpha^{fg}_{jk} = 1$ if class k is present in image j and $\alpha^{fg}_{jk} = 0$ otherwise. α^{bg} is always set to 1 to reflect the fact that background of different types can be shared across different images. That is, the foreground topics are clamped with the weak labels indicating the presence/absence of foreground object classes in each image, whilst all background types are assumed to be present a priori. With these partial constraints, iterating the updates in Eqs. (4)-(6) has the effect of factoring images into combinations of latent topics, where K^{bg} background topics are always available to explain away backgrounds, and K^{fg} foreground topics are only available to images with annotated classes. Note that this set-up assumes a 1:1 correspondence between object classes and topics. More topics can trivially be assigned to each object class (1:N correspondence), which has the effect of modelling multi-modality in object appearance, for a linear increase in computational cost.

Probabilistic feature fusion We combine multiple features probabilistically in our model. A single topic distribution (y) is estimated given different low-level features (f) in Eq. (5). Our fusion keeps the original low-level feature representations rather than increasing ambiguity by concatenating them before they provide complementary information about the location (early fusion). The shared

topic (y) and Gaussian location distribution ($\boldsymbol{\mu}, \Lambda^{-1}$) correlate the multiple features, which avoids domination by a single one. The appearance model in each modality is updated based on the consensus estimate of location; it thus learns a good appearance in each view even if the particular category is hard to detect in that view (as a result could drift if used alone). Its advantage over early (feature concatenation) or late (score level) fusion is demonstrated experimentally in Sec. 1 of the supplementary material.

5 Object Localisation

After learning, we extract the location of the objects in each image from the model, which can then be used to learn an object detector. Depending on whether the images are treated as individual images or consecutive video frames, our localisation method differs slightly.

Individual images There are two possible strategies to localise objects in individual images, which we will compare later in Sec. 8. In the first strategy (*Our-Gaussian*), a bounding box for class k in image j can be obtained directly from the Gaussian mean of the parameter posterior $q(\boldsymbol{\mu}_{jk}, \Lambda_{jk})$, via aligning a bounding box to the two standard deviation ellipse. This has the advantage of being clean and highly efficient. However, since there is only one Gaussian per class (which will grow to cover all instances of the class in an image), this is not ideal for images with more than one object per class. In the second strategy (*Our-Sampling*) we draw a heat-map for class k by projecting $q(y_{ijk})$ (Eq. (5)) back onto the image plane, using the grid coordinates where visual words are computed. This heat-map is analogous to those produced by many other approaches such as Hough transforms [65]. Thereafter, any strategy for heat-map based localisation may be used. We choose the non-maximum suppression (NMS) strategy of [5].

Video frames The above two strategies are directly applicable to video data if we treat each frame as an individual image. However, the temporal information of objects is useful in continuous videos to smooth the noise of individual frames. To this end, we apply a simple state space model for video segments to post-process object locations, smoothing them in time. Two diagonal points are sufficient to encode object location (bounding-box), and these are estimated from $q(\boldsymbol{\mu}_{jk}, \Lambda_{jk})$ above at every frame/time t as c_t. Assuming a four-dimensional state latent state vector $\mathbf{z}_t^T = (z_{xt}\ z_{yt}\ \dot{z}_{xt}\ \dot{z}_{yt})$, denoting the (hidden) true coordinate of an object of interest (two diagonal corners of the bounding box). A Kalman smoother is then adopted to smooth the observation noise σ_t in the system:

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \epsilon_t,\ \mathbf{c}_t = \mathbf{O}\mathbf{z}_t + \sigma_t, \tag{7}$$

where $\mathbf{A}$ is the temporal transition between true locations $\mathbf{z}$ in video, and $\mathbf{O}$ is the observation function for each frame.

6 Bayesian Priors

An important capability of our Bayesian approach is that top-down cues from human expertise, or estimated from data can be encoded. Various types of human knowledge about objects and their relationships with background are encoded in our model. As discussed earlier, prior cues can potentially cover appearance and geometry information.

Encoding geometry prior For geometry, we already model the most general intuition that objects are compact relative to background by assigning them Gaussian and uniform distributions respectively (Sec. 3). Beyond this, prior knowledge about typical image location and size of each class can be included via prior parameters $\boldsymbol{\mu}_k^0, \Lambda_k^0$, however we found this did not actually noticeably improve results in our experiments so we did not exploit it. This makes sense, because in challenging datasets like PASCAL VOC, objects appear in highly variable scales and locations, so there is little regularity to learn.

Encoding appearance prior If prior information is available about object category appearance, it can be included by setting $\boldsymbol{\pi}_{kf}^0$. (We will exploit this later for cross-domain adaptation in Sec. 7.1). For within-domain learning, we can obtain an initial data-driven estimate of object appearance to use as a prior by exploiting the observation that, when aggregated across all images, the background is more dominant than any single object class in terms of size (hence the amount of visual words). Exploiting this intuition, for each object class k, we set the appearance prior $\boldsymbol{\pi}_{kf}^0$ as:

$$\boldsymbol{\pi}_{kf}^0 = \left| \frac{1}{C} \sum_{j, c_j = k} h(\mathbf{x}_{jf}) - \frac{1}{J} \sum_{j} h(\mathbf{x}_{jf}) \right|_+ + \epsilon, \tag{8}$$

where $h(\cdot)$ indicates histogram and ϵ is a small constant. That is, set the appearance prior for each class to the mean of those images containing the object class minus the average over all images. This results in a prior which reflects what is consistently unique about each particular class. This is related to the notion of saliency, not within an image, but across all images. Saliency has been exploited in previous MIL based approaches to generate the instances/candidate object locations [11], [20], [12], [13], [21]. However, in our model it is cleanly integrated as a prior.

Encoding appearance similarity prior Going beyond the direct unary appearance prior discussed above, we next consider exploiting the notion of prior *inter-class appearance similarity*, rather than prior appearance per-se. The prior similarity between each object category can be estimated by computing inter-category category distance based on WordNet structure [33]. We compute a similarity matrix $\mathcal{M}$ where elements $\mathcal{M}_{m,n}$ indicates the relatedness between class m and n. The similarity matrix is then used to define how much appearance information from class m contributes to class n a priori.

We exploit this matrix by introducing an M-step into our learning algorithm (Eqs. (4)-(6)). Previously the appearance prior $\boldsymbol{\pi}_{kf}^0$ was considered fixed (e.g., from Eq. (8)). As with any parameter learning in the presence of latent variables, $\boldsymbol{\pi}_{kf}^0$ could potentially be optimised by a maximum-likelihood M-step interleaved with E-step latent variable inference. However, rather than the conventional approach of optimising $\boldsymbol{\pi}_{kf}^0$ *solely* given the data of class k, we define

an update that exploits cross-class similarity by updating $\boldsymbol{\pi}^0_{kf}$ using *all* the data, but weighted by its similarity to the target class k.

Denoting $\hat{\pi}^0_{vkf}$ as the new appearance prior to be learned, we introduce a new regularised M-step to learn $\hat{\pi}^0_{vkf}$. Specifically, the update for each class $k \in T^{fg}$ is as follows:

$$\hat{\pi}^0_{vkf} = \underbrace{\pi^0_{vkf}}_{\text{fixed data driven prior}} + \underbrace{\sum_{ij} \sum_{k' \in T^{fg}} \mathcal{M}_{k,k'} \cdot \mathbf{I}(x_{ijf} = v) \tilde{y}_{ijk'}}_{\text{inter-class similarity prior}} \tag{9}$$

The first term π^0_{vkf} is the original unary prior from Eq. (8). The second term is a data-driven update given the results of the E-step ($\tilde{y}$, Eqs. (4)-(6)). It includes a contribution from all images of all classes k', weighted by the similarity of k' to the target class k – given by $\mathcal{M}_{k,k'}$. The updated $\hat{\boldsymbol{\pi}}^0_{kf}$ then replaces $\boldsymbol{\pi}^0_{kf}$ in Eq. (6) of the E-step.

7 Learning from additional data

In this section, we discuss learning from additional data beyond the data for the WSOL task. This includes partially relevant data from other domains or datasets, and any additional but un-annotated data from the same domain.

7.1 Bayesian Domain Adaptation

Across different datasets or domains (such as images and video), the appearance of each object category will exhibit similarity, but vary sufficiently that directly using an appearance model learned in a source domain s for inference in a target domain t will perform poorly [18]. In our case this would correspond to directly applying a learned source appearance model π^s_k to a new target domain t, $\boldsymbol{\pi}^t_k := \boldsymbol{\pi}^s_k$. However, one hopes to be able to exploit similarities between the domains to learn a better model than using only the target domain alone [36], [37], [28], [40]. In our case, the Bayesian (Multinomial-Dirichlet conjugate) form of our model is able to achieve this for WSOL by simply learning $\boldsymbol{\pi}^s_k$ for a source domain s (Eq. (6)), and applying it as the prior $\boldsymbol{\pi}^{0^t}_k := \boldsymbol{\pi}^s_k$ in the target t – which is then adapted to reflect the target domain statistics (Eq. (6)).

7.2 Semi-supervised learning (SSL)

Beyond learning from annotated data in different but related domains, our framework can also be applied in a SSL context to learn from unlabelled data in the same domain to improve performance and/or reduce annotation requirement. Specifically, images j with known annotations are encoded as described in Sec. 4, while those without annotation are set to $\alpha^{fg}_j = 0.1 \ \forall j$, meaning that all topics/classes may potentially occur, but we expect few simultaneously within one image. Unknown images can include those from the same pool of classes but without annotation (for which the posterior $q(\boldsymbol{\theta})$ will pick out the present classes), or those from a completely disjoint pool of classes (for which $q(\boldsymbol{\theta})$ will encode only background).

8 Experiments

8.1 Datasets, features and settings

Datasets We evaluate our model on three datasets, PASCAL VOC [9], ImageNet [61] and YouTube-object video [17]. The challenging PASCAL VOC 2007 dataset is now widely used for weakly supervised object localisation. A number of variants are used: *VOC07-20* contains all 20 classes from VOC 2007 training set as defined in [20] and was used in [20], [12], [21]; *VOC07-6×2* contains 6 classes with Left and Right poses considered as separate giving 12 classes in total and was used in [11], [13], [20], [12], [21], [15]. The former obviously is more challenging than the latter. Note that *VOC07-20* is different to the *Pascal07-all* defined in [11] which actually contains 14 classes and uses the other 6 as fully annotated auxiliary data. We call it *VOC07-14* for consistency, but do not use the other 6 auxiliary classes.

To evaluate our method in a larger-scale setting, we select all images with bounding box annotation in the ImageNet dataset containing 3624 object categories as in [15].

We also evaluate our model on videos although it is designed primarily for individual images and does not exploit motion information during learning. Only a simple temporal smoothing post-processing step is introduced (see Sec. 5). YouTube-Object dataset [17] is a weakly annotated dataset composed of 10 object classes in videos from YouTube. These 10 classes are a subset of the 20 VOC classes, which facilitate domain transfer experiments.

Features By default, we use only a single appearance feature, namely SIFT to compare directly with most prior WSOL work which uses the same feature. Given an image j, we compute N_j 128-bin SIFT descriptors, regularly sampled every 5 pixels along both directions, and quantise them into a 2000-word codebook using K-means clustering. Differently to other bag-of-words (BoW) approaches [49], [50] which then discard spatial information entirely, we then represent each image j by the list of N_j visual words and corresponding locations $\{x_i, l_{ai}, l_{bi}\}_{i=1}^{N_j}$ where $\{l_{ai}, l_{bi}\}$ are the coordinates of each word.

We additionally extract two more BoW features at the same regular grid locations to test the feature fusion performance. They are: (1) Colour-LAB: Colour provides complementary information to SIFT gradients. We quantise colour histograms into three channels (8,16,16) of LAB space and concatenate them to produce a 40 dimensional feature vector. Visual words are then obtained by quantising the feature space using K-means with K=500. (2) Local binary pattern (LBP) [66]: 52 bin LBP feature vectors are computed and quantised into a 500-bin histogram.

Settings and implementation details For our model, we set the foreground topic number K^{fg} to be equal to the number of classes, and $K^{bg} = 20$ for background topics. α is set to 0 or 1 as discussed in Sec. 4. and π^0 is initialised by Eq. (8) as described in Sec. 5. μ^0 is initialised with the central of the image area. Λ^0 is initialised from the half size of the image area. We run Eqs. (4)-(6) for a fixed 100 VMP

iterations. The localisation performance is measured using CorLoc [17], [15]: an object is considered to be correctly localised in an given image if the overlap between the localisation box and the ground-truth (any instance of the target class) is greater than 50%. The CorLoc accuracy is then computed as the percentage (%) of correctly localised images for each target class. The same measure has been used in all methods compared in our experiments.

8.2 Comparison with state-of-the-art

8.2.1 Results on VOC dataset

Competitors We compare our joint modelling approach to the following state-of-the-art competitors:

Deselaers et al *[11]* A CRF-based multi-instance approach that localises object instances while learning object appearance. They report performance both with a single feature (GIST) and four appearance features (GIST, colour histogram, BoW of SURF, and HOG).

Pandey and Lazebnik [13] They adapt the fully supervised deformable part-based models to address the weakly supervised localisation problem.

Siva and Xiang [20] A greedy search method based on Genetic Algorithm to localise the optimal object bounding box location against a costing function combining the object saliency, intra-class and inter-class cues.

Siva et al *NM [12]* A simple negative mining (NM) approach which shows that inter-class is a stronger cue than the intra-class one when used properly.

Siva et al *OS [60]* The negative mining approach above is extended to mine objective saliency (OS) information from a large corpus of unlabelled image. This can be considered as a hybrid of the object saliency approach in [27] and the negative mining work in [12].

Shi et al *[21]* A ranking based transfer learning approach using an auxiliary dataset to score each candidate bounding box location in an image according to the degree of overlap with the unknown true location.

Zhu et al *[67]* An unsupervised saliency guided approach to localise an object in a weakly labelled image in a multiple instance learning framework.

Tang et al *[15]* An optimisation-centric approach that uses a convex relaxation of the MIL formulation.

Note that a number of the competitors [11], [21], [20], [12], [15] used an additional auxiliary dataset that we do not use. Objectness trained on auxiliary data was required by [11], [21], [20], [12], [15]. Although Shi et al. [21] evaluated all 20 classes, a randomly selected 10 were used as auxiliary data with bounding-boxes annotation. Pandey and Lazebnik [13] set aspect ratio manually and/or performed cropping on the obtained bounding-boxes.

Initial localisation Table 2 shows that for the initial annotation accuracy our model consistently outperforms all competitors over all three VOC variants, sometimes by big margins. This is mainly due to the unique joint modelling approach taken by our method, and its ability to integrate prior spatial and appearance knowledge in a principled way. Note that the prior knowledge is either based on first principle (spatial and appearance) or computed from the data without any additional human intervention (appearance). Our two object localisation methods (Our-Sampling and Our-Gaussian) vary in performance over different-sized datasets. Our-Gaussian performs better in the relatively simple datasets (6×2) where most images contain only one object, because our Gaussian location model can compact objects easily in this case. In contrast, Our-Sampling is better in the more complicated situation (20 classes) where many objects co-existing in one image is more common.

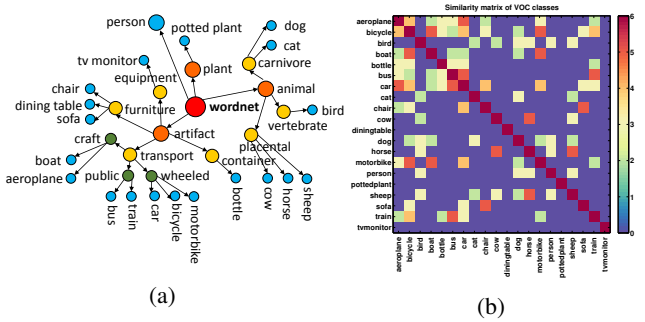
Method	Initialisation			Refined by detector		
	6×2	*14*	*20*	*6×2*	*14*	*20*
Deselaers *et al* [11]						
a. single feature	35	21	-	40	24	-
b. all four features	39	22	-	50	28	-
Pandey and Lazebnik [13] *						
a. before cropping	36.7	20.0	-	59.3	29.0	-
b. after cropping	43.7	23.0	-	61.1	30.3	-
Siva and Xiang [20]	40	-	28.9	49	-	30.4
Siva *et al* NM [12]	37.1	-	29.0	46	-	-
Siva *et al* OS [60]	42.4	-	31.1	55	-	32.0
Shi *et al* [21] +	39.7	-	32.1	-	-	-
Zhu *et al* [67]	-	-	-	-	31	-
Tang *et al* [15]	39	-	-	-	-	-
Cinbis *et al* [68]		-	-	-	-	**38.8**
Our-Sampling	50.8	**32.2**	**34.1**	65.5	**33.8**	36.2
Our-Gaussian	**51.5**	30.5	31.2	**66.1**	32.5	33.4
Our-Sampling+prior	51.2	**33.4**	**36.1**	65.9	**35.4**	38.3
Our-Gaussian+prior	**51.8**	31.1	33.5	**66.7**	33.0	35.8

TABLE 2: Comparison with state-of-the-art competitors on the three variations of the PASCAL VOC 2007 dataset. * Requires aspect ratio to be set manually. + Require 10 out of the 20 classes fully annotated with bounding-boxes and used as auxiliary data.

Refined by detector After the initial annotation of the weakly labelled images, a conventional strong object detector can be trained using these annotations as ground truth. The trained detector can then be used to iteratively refine the object location. We follow [13], [20] in exploiting a deformable part-based model (DPM) detector[1] [5] for one iteration to refine the initial annotation. Table 2 shows that again our model outperforms almost all competitors by a clear margin for all three datasets (see the supplementary material for more detailed per-class comparisons). Very recently, [68] achieved similar performance by training a multi-instance SVM with a more powerful fisher vector based representation.

With appearance similarity prior As described before, the proposed framework can exploit the appearance similarity prior across classes. Although the actual appearance similarity between classes is hard to calculate, we can approximate it by computing the relatedness using WordNet semantic tree [69]. Fig. 5 shows the pairwise relatedness among 20 classes, which is generated using the Lin distance of [33]. The diagonal of the matrix verifies that classes are most similar to themselves. Leaf nodes (blue) correspond to the classes of VOC-20. Classes that inherit from the same subtree should show more similar appearance. A

1. Version 3.0 is used for fair comparison against most published results obtained using the same version.



(a) (b)

Fig. 3: (a) A hierarchical structure of the 20 PASCAL VOC classes using WordNet. (b) The class similarity matrix.

pairwise similarity matrix is then calculated from the tree structure and used to correlate their appearance as explained in Sec. 5. The bottom two rows of Table 2 show the localisation accuracy with the appearance similarity prior. It clearly shows that the prior improves the performance of both variants of our model for all experiments. It is interesting to note that the performance is improved more on VOC-20 than VOC-6×2. This is because there is more opportunity to share related appearance as the number of classes increases. Categories in 6×2 are generally more dissimilar, so there is less benefit to the correlation.

What has been learned Fig. 4 gives examples of the localisation results and illustrates what has been learned for the foreground object classes. For the latter, we show the response of each learned object topic (i.e. the posterior probability of the topic given the visual word) as a gray-level image, or heat map (the brighter, the higher probability that the object is present at each image location). These examples show that the foreground topics indeed capture what each object class looks like and can distinguish it from the background and between different object classes. For instance, Fig. 4(c) shows that the motorbike heat map is quite accurately selective, with minimal response obtained on the other vehicular clutter. Fig. 4(e) indicates how the Gaussian can sometimes give a better bounding box. The opposite is observed in Fig. 4(f) where the single Gaussian assumption is not ideal when the foreground topic has less a compact response. Selectivity is illustrated by Fig. 4(c,d), Fig. 4(h,i) and Fig. 4(g,k), which show the same images, but with detection results for different co-occurring objects. In each case, the relevant object has been successfully selected while "explaining away" the potentially distracting alternative. Our method may fail if the background clutter or objects of no interest dominates the image (Fig. 4(l,m,u)). For example, in Fig. 4(l), a bridge structure resembles the boat in Fig. 4(a) resulting strong response from the boat topic, whilst the actual boat, although picked up, is small and overwhelmed by the false response.

A key strength of our framework is explicit modelling of background without any supervision. This allows background pixels to be explained, reducing confusion with foreground objects and hence improving localisation accuracy. This is illustrated in Fig. 5 via plots of the background

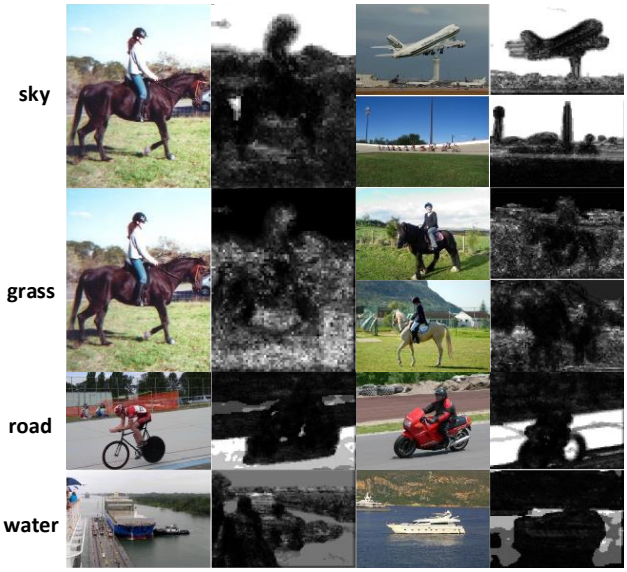


Fig. 5: Illustration of the learned background topics.

topic response (heat map). It illustrates qualitatively that some background topics are often correlated with common semantic background components such as sky, grass, road and water, despite none of these being annotated.

Weakly supervised detector The ultimate goal of weakly supervised object localisation is to learn a weakly supervised detector. This is achieved by feeding the localised objects into an off-the-shelf detector training model. The deformable part based model (DPM) in [5] is used and this weakly supervised (WS) detector is compared against a fully supervised (FS) one with the same DPM model (version 3.0). Specifically, Table 3 compares the mean average precision (mAP) of detection performance on both VOC-6×2 and VOC-20 test datasets among previous reported WS detector results, ours and the fully supervised detector [5]. Due to the better localisation performance on the weakly supervised training images, our approach is able to reduce the gap between the WS detector and the FS detector. The detailed per-class result is included in the supplementary material and it shows that for classes with high localisation accuracy (e.g. bicycle, car, motorbike, train), the WS detector is often as good as the the FS one, whilst for those with very low localisation accuracy (e.g. bottle and pottedplant), the WS detector fails completely.

Method	Deselaers [11]	Pandey [13]	Siva [20]	**Ours**	Fully Supervised
6×2	21	20.8	-	26.1	33.0
20	-	-	13.9	17.2	26.3

TABLE 3: Performance of strong detectors trained using annotations obtained by different WSOL methods

8.2.2 Results on ImageNet dataset

Table 4 shows the initial annotation accuracy of different methods for the much larger 3624-class ImageNet dataset.

Fig. 4: Top row in each subfigure: examples of object localisation using our-sampling and our-Gaussian. Bottom row: illustration of what is learned by the object (foreground) topics via heat map (brighter means object is more likely). The first four rows show some examples of PASCAL VOC and last two rows are selected from ImageNet.

Method	Initialisation
Alexe *et al* [27]	37.4
Tang *et al* [15]	53.2
Our-Sampling	**57.6**

TABLE 4: Initial annotation accuracy on ImageNet dataset

Note that the result of Alexe *et al* [27] is taken from the Table 4 in [15]. Although the annotation accuracy could be further improved by training an object detector to refine the annotation as shown in Table 2, this step is omitted in our experiment as none of the competitors attempted it. For such a large scale learning problem, loading all the image features into the memory is a challenge for our joint learning method. A standard solution is taken, that is, to process in batches of 100 classes. Joint learning is performed within each batch but not across batches; our model is thus not used to its full potential. Table 4 shows that our method achieves the best result (57.6%). Note that [27] is a very simple baseline as it simply takes the top-scoring objectness box. Recently more sophisticated transfer-based techniques [10] and [70] were evaluated on ImageNet. But their results were obtained on a different subset of ImageNet, thus not directly comparable here.

To investigate the effect of the similarity prior in this larger dataset, we randomly choose 500 small (containing around 100 images each) leaf-node classes from ImageNet for joint-learning with an inter-class similarity prior. This was the largest dataset size that could simultaneously fit in the memory of our platform[2]. Performing joint learning with inter-class correlation on this ImageNet subset, we achieve 58.8% annotation accuracy on the 500 classes

2. Our learning algorithm could potentially be modified to process all 3624 classes in batches.

compared to 55.4% without using the similarity prior.

8.2.3 Results on YouTube-object dataset

Our main competitors on YouTube-Object (YTO) are [17] and [23]. Prest *et al* [17] first performed spatio-temporal segmentation of video into a set of 3D tubes, and subsequently searched for the best object location. Very recently, [23] simultaneously localised objects of the same class across a set of video clips (co-localisation) with the Frank-Wolfe Algorithm. Note that there are some recently published studies on weakly supervised object segmentation from video [42]. This is not directly comparable as they did not report results based on the standard YTO bounding-box annotations. Two variants of our model are compared here: Our-sampling is the method evaluated above for individual images. Used here, it ignores the temporal continuity of the video frames in a video. Our-smooth is the simple extension of our sampling for video object localisation. As described in Sec. 5, temporal information is used to enforce a smooth change of object location over consecutive frames. The way temporal information is exploited is thus much less elaborative than that in [17]. For all methods compared, We evaluate localisation performance on the key frames which are provided with ground truth labels by [17].

Table 5 shows that even without using any temporal information and operating on key frames only, Our-sampling outperforms the method in [17]. Our-Smooth further improves the performance and the localisation accuracy of 32.2% is very close to the upper bound result (34.8%) suggested by [17], which is the best possible result from oracle tube extraction. Fig. 6 shows some examples of video object localisation using Our-Smooth. We note that all these results have been exceeded (50.1% accuracy) recently by a model purposefully designed for video segmentation [71], which performed much more intensive spatio-temporal modelling and used superpixel segmentation within each frame and motion segmentation across frames.

Categories	[17]	[23]	Our-Sampling	Our-Smooth	[71]
aeroplane	51.7	27.5	40.6	45.9	65.4
bird	17.5	33.3	39.8	40.6	67.3
boat	34.4	27.8	33.3	36.4	38.9
car	34.7	34.1	34.1	33.9	65.2
cat	22.3	42.0	35.3	35.3	46.3
cow	17.9	28.4	18.9	22.1	40.2
dog	13.5	35.7	27.0	27.2	65.3
horse	26.7	35.6	21.9	25.2	48.4
motorbike	41.2	22.0	17.6	20.0	39.0
train	25.0	25.0	32.6	35.8	25.0
Average	28.5	31.1	30.1	32.2	50.1

TABLE 5: Performance comparison on YouTube-object

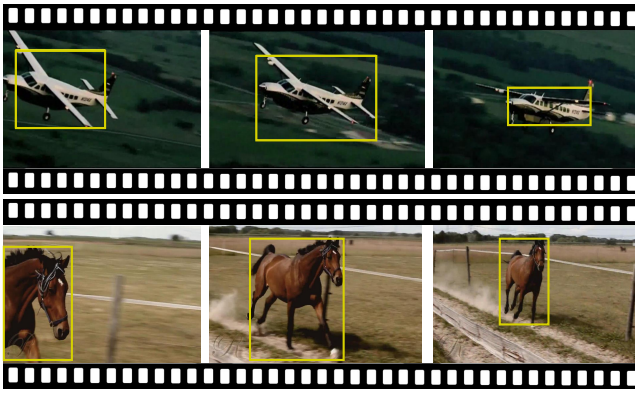


Fig. 6: Examples of video object localisation

8.3 Bayesian domain adaptation

We next evaluate the potential of our model for weakly supervised cross-domain transfer learning using the YouTube-Object and VOC07-10 as the two domains (we choose the same 10 classes from the VOC07-20 as in YouTube-Object). One domain contains continuous and highly varying video data, and the other contains high resolution but cluttered still images. We consider following two non-transfer baselines:

YTO, VOC The first baseline is the original performance on YouTube-Object and VOC07-10 classes, solely using target domain data. *YTO* is exactly the same as Our-Sampling described in Sec. 8.2.3, while *VOC* is trained with 10 classes from VOC07-20 using the same setting described in Sec. 8.2.1.

All→YTO, All→VOC The second baseline simply combines the training data of YouTube-Object and VOC. One model trained with these two domains' data is used to localise object on YouTube-Object ($A \rightarrow Y$) and VOC07-10 ($A \rightarrow V$).

We consider two directions of knowledge transfer between YouTube-Object and VOC07-10, and compare the above baselines with our domain adaptation method: $V \rightarrow Y$ is initialised with an appearance prior transferred from VOC07-10, and adapted on the YTO data. On the contrary, $Y \rightarrow V$ adapts the YTO appearance prior to VOC07-10. Table 6 shows that our Bayesian domain adaption method performs better than the baselines on both YouTube-Object and VOC07-10. In contrast, the standard combination (A→Y and A→V) shows little advantage over solely using target domain data. Note that unlike prior studies of video→image [17] or image→video [28] that adapt detectors with fully labelled data, our task is to adapt weakly labelled data.

We also vary the amount of target domain data and evaluate its effect on the domain transfer performance. Fig. 7 shows that our model provides a bigger margin of benefit given less target domain data. This can be easily understood because with a small quantity of training examples there is insufficient data to learn the object appearance well and the impact of the knowledge transfer is thus more significant.

8.4 Semi-supervised Learning

One important advantage of our model is the ability to utilise unlabelled data to further reduce the manual annotation requirements. To demonstrate this we randomly select 10% of the *VOC07-6×2* data as our weakly labelled training data, and then vary the additional unlabelled data used. Note that 10% labelled data corresponds to around

Categories	YTO			VOC		
	Y	A→Y	V→Y	V	A→V	Y→V
aeroplane	40.6	40.8	**45.8**	57.5	58.1	**58.7**
bird	39.8	**40.3**	38.8	29.8	30.5	**33.7**
boat	33.3	33.4	**38.8**	28.0	27.9	**29.0**
car	**34.1**	33.9	33.6	39.1	39.1	**44.4**
cat	35.3	35.3	**38.8**	59.0	**59.3**	58.6
cow	18.9	19.0	**27.7**	36.7	36.9	**38.9**
dog	27.0	**27.1**	26.7	46.5	47.4	**48.3**
horse	21.9	22.1	**26.1**	53.2	53.5	**55.5**
motorbike	17.6	**17.9**	17.5	55.6	55.2	**58.1**
train	32.6	32.6	**36.2**	54.7	54.5	**56.3**
Average	30.1	30.2	**33.0**	46.0	46.2	**48.1**

TABLE 6: Cross-domain transfer learning results

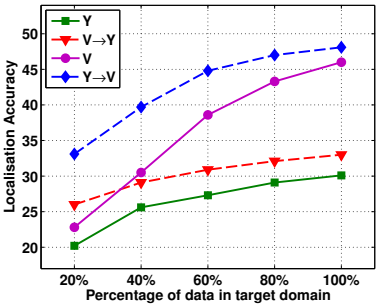


Fig. 7: Domain adaptation provides more benefit with fewer target domain samples.

only *5 weakly labelled images per class* for the *VOC07-6×2* dataset, which is significantly less than what any previous method has exploited. Two evaluation procedures are considered: (i) Evaluating localisation performance on the initially annotated 10% (standard WSOL task); and (ii) WSOL performance on the held out *VOC07-6×2* test set[3]. The latter corresponds to an online application scenario where the localisation model is trained on one database and needs to be applied online to localise objects in incoming weakly labelled images. We vary the additional data across a combination of four conditions: (1) $6R$: add the remaining 90% of data for the 6 target classes but without labels, (2) $100U$: add all images from 100 unrelated ImageNet classes without labels, (3) $6R+100U$: add both of the above. There are two questions to answer: Whether the model can exploit the related data when it comes without labels (6R), and whether it can avoid being confused by a vast quantity of unrelated data (100U).

The results are shown in Table 7, where the ratio of relevant to irrelevant data in the additional unlabelled samples is shown in the second column. From the results, we can draw the following conclusions: (1) As expected, the model performs poorly with little data (10%L). However it improves significantly with some relevant but unlabelled data (the standard SSL setting, 10%L+6R). Moreover, this SSL result is almost as good as when all the data is labelled (100%L). (2) If *only* irrelevant data is added to the small labelled seed, not only does the performance not degrade, but it increases noticeably (10%L vs. 10%L+100U). (3) If both relevant and irrelevant data are added – corresponding to the realistic scenario where an automatic process gathers a pool of potentially relevant data which, without any screening, will be a mix of relevant and irrelevant data to the target problem. In this case the performance improves to not far off the fully annotated case (10%L vs. 10%L+6R+100U vs. 100%L). As expected, the performance of 10%L+6R+100U is weaker than 10%L+6R – if one manually goes through the unlabelled data and removes the irrelevant ones and leave only the relevant ones, it would certainly benefit the model. But it is noted that the decrease in performance is small (47.1% to 43.5%). (4) If the irrelevant data is added to the fully annotated dataset, the performance improves slightly (100%L vs. 100%L+100U), which shows that our model is robust to this potential distraction from the large amount of unlabelled and irrelevant data. This is expected in SSL, which typically benefits only when the amount of labelled data is small. These results show that our approach has good promise for effective use in realistic scenarios of learning from only few weak annotations and a large volume of only partially relevant unlabelled data. This is illustrated visually in Fig. 8, where unlabelled data helps to learn a better object model. Finally, the similarly good results on the held-out test set verify that our model is indeed learning a good generalisable localisation mechanism and is not over-fitted to the training data.

3. To localise objects in a test image, we only need to iterate Eqs. (4)-(5) instead of (4)-(6). That is, the object appearance is considered fixed and does not need to be updated. This both reduces the cost of each iteration and also makes convergence more rapid.

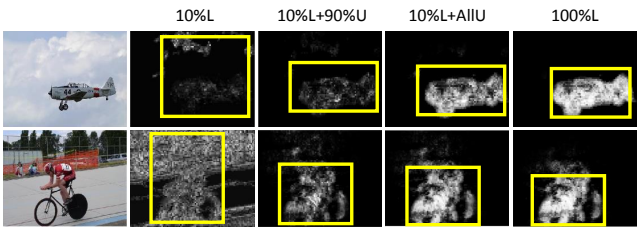


Fig. 8: Unlabelled data improves foreground heat maps.

VOC07-6 × 2		Data for Localisation	
Data for Training	ratio of R:U	*10%L*	*Test set*
10%L	-	27.1	28.0
10%L+6R	1	47.1	42.3
10%L+100U	0	35.8	32.4
10%L+6R+100U	0.04	43.5	38.1
100%L	-	50.3	46.2
100%L+100U	0	50.7	47.5

TABLE 7: Localisation performance of semi-supervised learning using *Our-Sampling*

8.5 Computational cost

Our model is efficient both in learning and inference, with a complexity $\mathcal{O}(NMK)$ for N images, M observations (visual words) per image, and K classes. The experiments were done on a 2.6GHz PC with a single-threaded Matlab implementation. Training the model on all 5,011 VOC07

images required 3 hours and a peak of 6 GB of memory to learn a joint model for 20 classes. Our Bayesian topic inference process not only enables prior knowledge to be used, but also achieves 10-fold improvements in convergence time compared to EM inference used by most conventional topic models with point-estimated Dirichlet topics. Online inference of a new test image took about 0.5 seconds. After model learning, for object localisation in training images, direct Gaussian localisation is effectively free and heat-map sampling took around 0.6 seconds per image. These statistics compare favourably to alternatives: [11] reported 2 hours to train 100 images; while our Matlab implementations of [12], [20] and [52] took 10, 15 and 20 hours respectively to localise objects for all 5,011 images.

9 Conclusion and Future work

We have presented an effective and efficient model for weakly-supervised object localisation (WSOL). Our approach surpasses the performance of prior methods and obtains state-of-the-art results on PASCAL VOC 2007 and ImageNet datasets. It can also be applied to the YouTube-Object dataset, and to domain transfer between these image and video datasets. With joint multi-label modelling, instead of independent learning in previous work, our model enables: (1) exploiting multiple object co-existence within images, (2) learning a single background shared across classes and (3) dealing with large scale data more efficiently than prior approaches. Our generative Bayesian formulation, enables a number of novel features: (1) integrating appearance and geometry priors, (2) exploiting inter-category appearance similarity and (3) exploiting different but related datasets via domain adaptation. Furthermore, it is able to use (potentially easier to obtain) unlabelled data with a challenging mix of relevant and irrelevant images to obtain an reasonable localiser when labelled data are in short supply for the target classes.

In this study we showed the usefulness of top-down, cross-class and domain transfer priors – demonstrating the model's potential to scale learning through transfer [21], [10], [7]. These contributions bring us significantly closer to the goal of scalable learning of strong models from weakly-annotated non-purpose collected data on the Internet.

It is worth pointing out that apart from adding a few new features (e.g. foreground-background topic separation and effective supervision via topic clamping), our generative Bayesian topic model is not fundamentally different from existing topic models used for image understanding [49], [45]. Nevertheless, state-of-the-art WSOL performance is obtained compared with more popular, more highly engineered and complex, and slower discriminative models. This not only shows the importance of the change of paradigm from independent discriminative learning to joint generative learning, but also suggests that sometimes it is not necessary to invent a completely new model; finding the missing ingredients that make an existing model work can be equally important.

Possible directions for future work include: automatically determining the optimal number of topics K [56], learning a deeper multi-layered [56] model by exploiting parts [22], [56] and attributes [72] rather than the current flat model; learning rather than pre-defining object-appearance similarity [16]; and learning from realistically noisy non-purpose collected labels [72].

References

[1] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *ECCV*, 2010.

[2] T. Hospedales, J. Li, S. Gong, and T. Xiang, "Identifying rare and subtle behaviors: A weakly supervised joint topic model," *TPAMI*, 2011.

[3] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box," in *ICCV*, 2009.

[4] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in *ICCV*, 2009.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, 2010.

[6] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu, "Multiple Component Learning for Object Detection," in *ECCV*, 2008.

[7] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *ECCV*, 2012.

[8] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," *CVPR*, 2013.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[10] M. Guillaumin and V. Ferrari, " Large-scale Knowledge Transfer for Object Localization in ImageNet," in *CVPR*, 2012.

[11] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *IJCV*, 2012.

[12] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *ECCV*, 2012.

[13] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011.

[14] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, "Weakly supervised object recognition and localization with stable segmentations," in *ECCV*, 2008.

[15] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *CVPR*, 2014.

[16] R. Salakhutdinov, A. Torralba, and J. Tenenbaum, "Learning to share visual appearance for multiclass object detection," in *CVPR*, 2011.

[17] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *CVPR*, 2012.

[18] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[19] X. Zhu, "Semi-supervised learning literature survey." University of Wisconsin-Madison Department of Computer Science, Tech. Rep. 1530, 2007.

[20] P. Siva and T. Xiang., "Weakly supervised object detector learning with model drift detection," in *ICCV*, 2011.

[21] Z. Shi, P. Siva, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *BMVC*, 2012.

[22] D. Crandall and D. Huttenlocher, "Weakly supervised learning of part-based spatial models for visual object recognition," in *ECCV*, 2006.

[23] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-wolfe algorithm," in *ECCV*, 2014.

[24] O. Maron and T. Lozano-Prez, "A framework for multiple-instance learning," in *NIPS*, 1998.

[25] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2003.

[26] N. Nguyen, "A new svm approach to multi-instance multi-label learning," in *ICDM*, 2010, pp. 384–392.

[27] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, 2012.

[28] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *NIPS*, 2012.

[29] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *CVPR*, 2010.

[30] D. Kuettel and V. Ferrari, "Figure-ground segmentation by transferring window masks," in *CVPR*, 2012.

[31] A. Zweig and D. Weinshall, "Exploiting Object Hierarchy: Combining Models from Different Category Levels," in *ICCV*, 2007.

[32] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *CVPR*, 2011.

[33] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity: measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL 2004*, 2004.

[34] L. Cao, Z. Liu, and T. Huang, "Cross-dataset action detection," in *CVPR*, 2010.

[35] S. J. Pan and Q. Yang, "A survey on transfer learning," *TKDE*, 2010.

[36] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *ACM MM*, 2007.

[37] L. T. Alessandro Bergamo, "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach," in *NIPS*, 2010.

[38] L. Jie, T. Tommasi, and B. Caputo, "Multiclass transfer learning from unconstrained priors," in *ICCV*, 2011.

[39] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, "Semi-supervised domain adaptation with instance constraints," in *CVPR*, 2013.

[40] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Transferring naive bayes classifiers for text classification," in *AAAI*, 2007.

[41] M. B. Blaschko, A. Vedaldi, and A. Zisserman, "Simultaneous object detection and ranking with weak supervision," in *NIPS*, 2010.

[42] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *CVPR*, 2013.

[43] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar, "Weakly supervised learning of object segmentations from web-scale video," in *ECCV*, 2012.

[44] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, 2003.

[45] J. Philbin, J. Sivic, and A. Zisserman, "Geometric latent dirichlet allocation on a matching graph for large-scale image datasets," *IJCV*, 2011.

[46] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *ICCV*, 2005.

[47] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *TPAMI*, 2006.

[48] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent object segmentation and classification." in *ICCV*, 2007.

[49] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding:classification, annotation and segmentation in an automatic framework," in *CVPR*, 2009.

[50] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *CVPR*, 2009.

[51] N. Rasiwasia and N. Vasconcelos, "Latent dirichlet allocation models for image classification," *TPAMI*, 2013.

[52] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR*, 2003.

[53] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *NIPS*, 2011.

[54] Z. Zhou and M. Zhang, "Multi-instance multilabel learning with application to scene classification," in *NIPS*, 2007.

[55] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *CVPR*, 2008.

[56] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed objects and parts," *IJCV*, 2008.

[57] P. V. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *ICCV*, 2009.

[58] P. Gehler and S. Nowozin, "Let the kernel figure it out; principled learning of pre-processing for kernel classifiers," in *CVPR*, 2009.

[59] F. Orabona, L. Jie, and B. Caputo, "Online-batch strongly convex multi kernel learning," in *CVPR*, 2010.

[60] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *CVPR*, June 2013.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[62] Z. Shi, T. M. Hospedales, and T. Xiang, "Bayesian joint topic modelling for weakly supervised object localisation," in *ICCV*, 2013.

[63] J. Winn and C. M. Bishop, "Variational message passing," *JMLR*, 2005.

[64] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[65] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *TPAMI*, 2011.

[66] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, 2002.

[67] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *TPAMI*, 2014.

[68] R. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold mil training for weakly supervised object localization," in *CVPR*, 2014.

[69] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

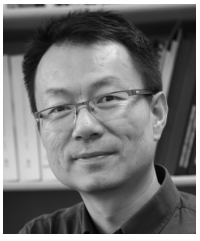
[70] A. Vezhnevets and V. Ferrari, "Associative embeddings for large-scale knowledge transfer with self-assessment," in *CVPR*, 2014.

[71] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *ICCV*, 2013.

[72] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *ECCV*, 2012.

Zhiyuan Shi received the BEng degree in electronic engineering and computer science from Beijing University of Posts and Telecommunications in 2011. He is currently a PhD student in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include weakly supervised learning, topic model, object localisation and attribute learning.

Timothy M. Hospedales received the PhD degree in neuroinformatics from the University of Edinburgh in 2008. He is currently a lecturer (assistant professor) of computer science at Queen Mary University of London. His research interests include probabilistic modelling and machine learning applied variously to problems in computer vision, data mining, interactive learning, and neuroscience. He has published more than 20 papers in major international journals and conferences. He is a member of the IEEE.



Tao Xiang received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 100 papers in international journals and conferences and co-authored a book, Visual Analysis of Behaviour: From Pixels to Semantics.